Prompt-Anchored Vision–Text Distillation for Lifelong Person Re-identification

Supplementary Material

Overview

This supplementary document provides additional ablation studies and evaluation protocols to complement the analysis in the main paper. We first report results on an alternative lifelong protocol where DukeMTMC-reID is replaced by LPW-s2, verifying the robustness of PAD under different domain compositions. We then examine the influence of distillation strength on both the textual and visual branches. Next, we provide additional ablations on AKA-order2 and on the number of unfrozen transformer blocks. Finally, we study the VA-Prompt pool allocation strategy under a fixed capacity. All experimental settings remain identical to those in the main paper unless otherwise specified.

A. Evaluation with LPW-s2

To validate the effectiveness on more evaluation protocols, we also report the results of PAD under the LPW-s2 [4] substitution protocol, where DukeMTMC-reID is replaced by LPW-s2 following recent practice in lifelong ReID. The rest of the configuration remains identical to the main paper. This evaluation is intended to verify that the behavior of PAD is stable when domain composition is slightly altered.

The performance on the LPW-based versions of AKA-order1 and AKA-order2 is summarized in Tabs. S1 and S2, where the final stage average mAP/R1 on seen and unseen domains is shown for both orders.

Across both orders, PAD retains the same qualitative behavior as reported in the main paper, and the results confirm that the proposed framework is robust to reasonable data changes in the lifelong sequence.

B. Ablation on Distillation Strength

B.1. Textual Distillation

The main paper applies a weak but stable textual distillation on top of the frozen text feature bank, aiming to keep the TA-Prompt semantically anchored while leaving sufficient room for domain-specific adaptation. To study the role of this component, we vary the overall strength of textual distillation by jointly modifying the KL weight λ_{logit}, temperature τ, logit scale γ, and the number of negatives, while keeping all other settings fixed.

We evaluate five configurations of textual distillation, denoted T1–T5, arranged from weak to strong. T1 adopts a lighter KL weight λ_{logit} and a mild logit scale γ. T2 corresponds to the baseline setting used in the main paper. T3 increases only λ_{logit} while keeping the temperature τ, logit scale γ, and the number of negatives per batch fixed, forming a clean "pure-weight" variant for isolating the effect of KL strength. T4 strengthens the distillation by lowering τ and substantially enlarging γ, and additionally increases the negative batch size to enlarge the contrastive space, producing a sharper and more globally stable teacher signal. T5 further increases λ_{logit} on top of this sharper configuration while slightly reducing γ relative to T4 (still higher than T2), and maintains the larger negative batch size, ensuring a monotonic progression of overall distillation strength. In other words, T4 emphasizes sharper logits through a larger γ, whereas T5 emphasizes a stronger KL constraint via a larger λ_{logit}, allowing the two settings to reinforce textual distillation along complementary dimensions. Results for both seen and unseen domains under AKA-order1 are summarized in Tab. S3, and all hyperparameters match the YAML blocks included in the main paper.

The results indicate that mild-to-moderate textual supervision (T1–T3) yields the most stable performance across stages, maintaining a consistent balance between seen and unseen domains. As the distillation becomes increasingly sharp (T4) or overly dominant (T5), the representation begins to bias toward the teacher distribution, leading to gradual degradation in both seen and unseen performance. These trends suggest that a weak textual constraint provides reliable trade-off between semantic alignment and continual adaptability.

B.2. Visual Distillation

PAD applies multi-level visual distillation through an EMA teacher, combining feature-level and logit-level alignment. The default configuration in the main paper uses a balanced weighting $(\lambda_{\text{feat}}, \lambda_{\text{logit}}) = (0.5, 0.5)$ with a temperature $\tau = 4.0$, while the EMA momentum is fixed to 0.997 throughout training.

To assess how the strength of visual supervision influences continual adaptation, we vary the three distillation parameters λ_{feat}, λ_{logit}, and τ while keeping all other components identical. Five variants are tested, ranging from very weak to very strong, constructed by progressively increasing the distillation weights and adjusting the temperature accordingly. Performance on both seen and unseen domains under AKA-order1 is summarized in Tab. S4.

The results show a clear upward trend from V1 to V3, indicating that moderate visual supervision effectively stabilizes the partially unfrozen backbone and yields consistent improvements across both seen and unseen domains. Once the distillation weights continue to increase and the temperature is lowered (V4–V5), the gains begin to saturate

Table S1. Performance comparison with LReID methods on **Training Order-1**: Market-1501 → CUHK-SYSU → LPW → MSMT17 → CUHK03. The optimal and suboptimal values are highlighted in red and blue.

Method	**Venue**	Market-1501		CUHK-SYSU		LPW		MSMT17		CUHK03		**Seen-Avg**		**Unseen-Avg**	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
PatchKD [1]	MM'22	**71.6**	**87.7**	77.0	79.6	33.2	41.9	7.0	18.5	29.5	30.4	43.7	51.6	47.8	41.4
LSTKC [3]	AAAI'24	57.0	78.6	82.9	84.9	47.2	58.4	**18.4**	**41.1**	42.3	43.7	49.6	61.3	57.8	50.2
DKP [2]	CVPR'24	60.0	80.3	**84.1**	**85.9**	46.0	57.9	17.7	38.5	41.0	41.4	49.8	60.8	57.5	**50.7**
SPRED [4]	ICCV'25	63.1	81.7	83.2	84.8	**50.6**	**60.7**	15.2	34.5	**48.6**	**50.0**	**52.1**	**62.3**	**58.7**	**50.7**
PAD (Ours)		**81.4**	**92.1**	**92.2**	**92.9**	**65.2**	**73.8**	**44.7**	**71.1**	**68.0**	**69.1**	**70.3**	**79.8**	**77.6**	**69.7**

Table S2. Performance comparison with LReID methods on **Training Order-2**: LPW → MSMT17 → Market-1501 → CUHK-SYSU → CUHK03. The optimal and suboptimal values are highlighted in red and blue.

Method	**Venue**	LPW		MSMT17		Market-1501		CUHK-SYSU		CUHK03		**Seen-Avg**		**Unseen-Avg**	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
PatchKD [1]	MM'22	**58.0**	**69.0**	6.3	16.7	46.3	70.6	75.7	78.5	29.6	30.2	43.2	53.0	45.3	38.5
LSTKC [3]	AAAI'24	46.7	57.6	**14.9**	**33.9**	56.5	78.0	84.0	86.1	42.1	43.7	48.8	59.9	57.4	49.5
DKP [2]	CVPR'24	49.5	61.4	14.1	32.6	**60.3**	**80.6**	**84.5**	**86.4**	**43.6**	43.7	**50.4**	**60.9**	**59.5**	**52.4**
SPRED [4]	ICCV'25	51.8	62.4	10.4	26.2	56.8	77.0	**84.5**	85.9	42.9	**43.8**	49.3	59.1	57.1	49.1
PAD (Ours)		**71.5**	**79.7**	**38.3**	**64.7**	**77.7**	**89.6**	**93.4**	**94.2**	**65.7**	**68.2**	**69.3**	**79.3**	**77.3**	**69.8**

Table S3. Effect of textual distillation strength (T1–T5). We report final stage average performance on both seen and unseen domains. Each configuration varies only in the KL weight λ_{logit}, temperature τ, and logit scale γ, whose values are listed in the table. T1–T3 use a negative batch size of 256, while T4–T5 increase it to 512 to enlarge the contrastive space. **T2** is used as the baseline in the main paper.

ID	λ	τ	γ	Seen		Unseen	
				mAP	R1	mAP	R1
T1	0.25	0.07	4.0	70.7	**81.3**	78.5	71.2
T2	**0.50**	**0.07**	**7.0**	70.7	81.0	**78.6**	**71.4**
T3	0.70	0.07	7.0	**70.8**	**81.3**	78.3	71.0
T4	0.70	0.05	16.0	70.3	81.0	78.2	71.1
T5	1.00	0.05	12.0	69.5	80.2	77.4	70.3

Table S4. Effect of visual distillation strength (V1–V5). We report final stage average performance on both seen and unseen domains. Each configuration varies only in the feature- and logit-level weights ($\lambda_{\text{feat}}, \lambda_{\text{logit}}$) and the temperature τ. The EMA momentum is fixed to 0.997 throughout training. **V3** corresponds to the configuration adopted in the main paper.

ID	λ_{feat}	λ_{logit}	τ	Seen		Unseen	
				mAP	R1	mAP	R1
V1	0.25	0.25	4.0	70.0	80.9	77.3	69.9
V2	0.35	0.35	4.0	70.2	**81.0**	77.7	70.5
V3	**0.50**	**0.50**	**4.0**	**70.7**	**81.0**	**78.6**	**71.4**
V4	0.75	0.75	3.5	70.5	80.9	78.2	70.8
V5	1.00	1.00	3.0	70.0	80.3	77.1	69.7

and eventually reverse, with stronger variants showing mild degradation in both metrics. This suggests that while a balanced level of visual guidance is beneficial, overly strong constraints still narrow the adaptation space and hinder the model's ability to adjust to new domains.

C. Component Analysis on AKA-order2

To verify that the component-wise conclusions in the main paper are not specific to AKA-order1, we further conduct the same ablation study on AKA-order2. Following the setting in the main paper, we report four configurations: VA-Prompt only (S2), S2 + TEXKD (S3), S2 + VISKD (S4), and the full PAD model (S5). The results are summarized in Table S5.

The trend is broadly consistent with AKA-order1. VA-Prompt already restores substantial plasticity, while TEXKD and VISKD further improve stability from the textual and visual sides. On AKA-order2, the gains of the two distillation pathways are not strictly monotonic, but the overall results still support their complementary roles under different lifelong orders.

D. Selective Layer Unfreezing

The main paper updates the last 4 transformer blocks of the visual backbone. To justify this choice, we compare four variants that unfreeze the last 2, 4, 6, or 8 blocks while keeping all other settings fixed.

Table S6 reports the final-stage seen-domain average, Market1501 performance as a proxy for early-domain re-

Table S5. Ablation study on AKA-order2. Columns indicate modules: **Freeze**—PAD freezing scheme, **VA**—Visual Adaptive Prompt, **TEXKD**—textual fixed distillation, **VISKD**—visual EMA distillation.

ID	Freeze	VA	TEXKD	VISKD	DukeMTMC		MSMT17		Market1501		CUHK-SYSU		CUHK03		Seen Avg	
					mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
S2	✓	✓			63.4	78.4	33.0	59.8	70.2	86.8	91.9	93.0	67.8	70.4	65.3	77.7
S3	✓	✓	✓		65.5	79.5	33.4	60.0	72.1	87.1	92.5	93.8	**67.9**	**70.7**	66.3	78.2
S4	✓	✓		✓	71.0	**82.7**	**38.0**	**64.5**	77.3	89.7	93.8	94.7	66.2	68.8	**69.3**	**80.1**
S5	✓	✓	✓	✓	**71.1**	82.6	37.7	64.0	**77.5**	**89.8**	**93.9**	**94.9**	66.4	68.5	**69.3**	80.0

Table S6. Effect of the number of unfrozen blocks. We report the final-stage seen-domain average, Market1501 performance, and trainable parameters. **4 blocks** corresponds to the configuration used in the main paper.

Blocks	Seen		Market1501		Trainable	
	mAP	R1	mAP	R1	Param	Ratio
2	68.7	79.3	**82.0**	91.8	23.6M	16.29%
4	70.7	81.0	81.2	**92.0**	37.8M	26.07%
6	71.9	82.3	77.5	90.0	52.0M	35.84%
8	**72.0**	**82.6**	74.3	88.3	66.2M	45.61%

tention, and the trainable parameter scale. Unfreezing more blocks slightly improves the overall average, but also reduces retention on the earliest domain and increases the training cost. Therefore, we adopt 4 blocks in the main paper as a practical balance between adaptation and efficiency.

E. Ablation on VA-Prompt Pool Allocation

The VA-Prompt module maintains a small pool of expert prompts, and the main paper adopts the allocation $[8, 8, 8, 8, 4]$, which reflects a combined consideration of dataset size and the ordering of domains. To assess the role of this allocation strategy, we keep the global prompts and the Top-K routing fixed, and vary only the distribution of expert slots.

We evaluate four alternative layouts: a uniformly small configuration with reduced overall capacity; a uniform-full design that preserves the total capacity of the default but distributes it evenly; a head-heavy variant that allocates more slots to earlier domains; and a tail-heavy version that emphasizes later domains. The final stage average performance on seen and unseen domains under AKA-order1 is summarized in Tab. S7.

The uniform-small configuration (P2) performs clearly worse than the default, showing that the expert pool cannot be compressed too aggressively without harming sequential performance. Among capacity-matched variants (P3–P5), the uniform-full layout remains slightly weaker than the default, indicating that a balanced but non-uniform assignment is more effective than an equal split. Both the head-heavy and tail-heavy allocations yield comparable but consistently lower results, suggesting that concentrating capacity at either end of the domain sequence provides no advantage. Overall, the results indicate that PAD is not overly sensitive to reasonable capacity-matched allocations, while overly compressed or strongly imbalanced assignments lead to inferior performance.

Table S7. Ablation on VA-Prompt pool allocation (P1–P5). We report final stage average performance on both seen and unseen domains. Each variant modifies only the allocation of expert prompt slots, while the global prompt remains unchanged. **P1** corresponds to the default configuration used in the main paper.

ID	Slot	Total	Seen		Unseen	
			mAP	R1	mAP	R1
P1	**[8,8,8,8,4]**	**36**	**70.7**	81.0	**78.6**	**71.4**
P2	[4,4,4,4,4]	20	68.7	80.0	76.5	69.0
P3	[8,7,7,7,7]	36	70.2	80.7	77.8	70.5
P4	[12,8,8,4,4]	36	70.5	**81.2**	77.7	70.2
P5	[4,4,8,8,12]	36	70.5	**81.2**	77.3	70.1

References

[1] Zhicheng Sun and Yadong Mu. Patch-based knowledge distillation for lifelong person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 696–707, 2022.

[2] Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16604–16613, 2024.

[3] Kunlun Xu, Xu Zou, and Jiahuan Zhou. Lstkc: Long short-term knowledge consolidation for lifelong person re-identification. In *AAAI*, 2024.

[4] Kunlun Xu, Fan Zhuo, Jiangmeng Li, Xu Zou, and Jiahuan Zhou. Self-reinforcing prototype evolution with dual-knowledge cooperation for semi-supervised lifelong person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.